图书情报工作

ChinaXiv合作期刊

第 64 卷 第 4 期 2020 年 2 月

大数据服务中知识组织的挑战及应对*

■ 张运良[1,2]

[1] 中国科学技术信息研究所 北京 100038 [2] 富媒体数字出版内容组织与知识服务重点实验室 北京 100038

摘 要：[目的/意义] 大数据服务的需求使得知识组织工作面临更大的挑战，通过发现、理解和分析这些挑战，把握知识组织工作的可能变化，提出应对方法。[方法/过程] 聚焦知识组织系统构建和应用，从大数据服务项目实践真实案例出发，分析知识组织不同角度的挑战，并提出应对策略。[结果/结论] 大数据服务中知识组织挑战可以分为数据膨胀、文献保证、集成和应用等 4 个方面，提出包含新型知识结构、多来源更新策略和弹性应用服务模式的面向大数据服务的系列知识组织框架，以便能够更好地应对上述挑战。

关键词：知识组织 大数据 知识服务 挑战 对策

分类号：G250

DOI：10.13266/j.issn.0252-3116.2020.04.010

大数据这一概念出现很早，但直到 2001 年数据分析师 D. Laney 定义了 3V 特性后才引起广泛关注和重视[1]，包括政府[2]、图书馆[3]、情报机构[4]等在内的多个行业均深受其影响。大数据服务是从大数据特点出发，调用大量分散数据和计算资源的服务，可以分为大数据查询服务和大数据分析服务两种方式[5]。查询和分析服务都离不开有组织的知识，对知识的加工、整理、揭示、控制等知识组织过程[6]有利于服务质量的提升[7]。大数据价值的充分发挥需要数据内部更为广泛的关联[8]，而这也是知识组织系统（Knowledge Organization System，KOS）的核心作用，新的环境下，知识组织系统本身也需要做出变革，以便应对学科以及数据本身变化并更具灵活性和适应性[1]，在大数据服务实践中，本体、开放链接数据、知识图谱等知识组织系统得到了较多的应用。

大数据知识组织工作得到了来自不同领域机构的关注和推动。高校、图书馆、情报所等研究服务机构注重理论探索和自身服务升级改造。研究机构在推动知识组织适应知识复用、发现和增值需要，将知识组织与计算技术相结合等方面已经开展探索，并在一些细分领域和行业初步取得了成效[9-11]。综合性服务机构更注重应用效果，如中国工程科技知识中心较早注意到知识组织在大数据服务中的作用，并大力推动知识组织和知识服务工作[12]。新闻出版机构注重从源头重塑信息资源，在原国家新闻出版广电总局推动下，国家知识资源服务中心开始建设，知识组织相关行业标准制定发布[13]，多家出版机构利用自有专业数字内容构建的特色知识服务开始上线应用。这些研究和实践为本研究提供了一定的基础。

1 知识组织挑战

1.1 数据膨胀挑战

知识组织工作的核心是通过规范实体以及实体之间关联进而通过标引规范数据，在词表、分类法、词系统、本体等不同类型的知识组织系统中，这些实体可能是概念、词条、类目等。大数据环境下实体数目增长迅速，如《汉语主题词表》1980 年出版时具有正式主题词 91 958 条，非正式主题词 17 410 条[14]，而 2014 年出版的《汉语主题词表（工程技术卷）》就收录优选词 19.6 万条，非优选词 16.4 万条；2018 年《汉语主题词表（自然科学卷）》则收录专业术语 12.4 万条[15]。清华大学构建的 XLORE 知识图谱则包含超过 1 628 万个实体[16]。

随着知识组织体量扩大，不同实体之间的关联也日益丰富。一般的叙词表中的关系仅包含用、代、属、分、参、族等，属性包括多语言（翻译）、定义、范围注

* 本文系中国工程科技知识中心项目“知识组织体系建设”（项目编号：CKCEST-2019-2-2）和中国科学技术信息研究所重点工作项目“科技创新大数据决策分析平台建设”（项目编号：ZD2019-08）研究成果之一。

作者简介：张运良（ORCID：0000-0003-4987-1539），研究员，博士，E-mail：zhangyl@istic.ac.cn。

收稿日期：2019-05-09 修回日期：2019-09-03 本文起止页码：88-94 本文责任编辑：杜杏叶

释、历史注释、一般注释等,一般都不超过 10 种,而美国国立医学图书馆汇集的通用医学语言系统(Unified Medical Language System,UMLS)的语义网络中就包含了 54 种语义关系[17];中国科学技术信息研究所的汉语科技词系统中语义关系类型更为丰富,在其新能源汽车卷中二级关系类型有 78 种,二级属性类型有 45 种;中国工程科技知识中心词表(核心集)有各类关系 399 种[18];Cyc 知识库中包含各类关系共 42 500 种[19];XLORE 知识图谱包含各类关系超过 44.6 万种[16]。

实体关联类型膨胀为知识组织系统的构建和应用带来了挑战。在手工构建实践中,知识工程师需要为两个实体确认一种关系,在数十种关系中选择所花费的时间要远远大于从数种关系中选择,而准确率和一致性则会下降。在自动构建实践中,由于很难处理同义、近义的冗余关联类型,加之实体众多,很难一一审核关联是否准确,知识组织系统的膨胀更为急遽。如在某个特定的语料库中对所有的词以某种特定的相关度计算方法,根据不同的阈值判断该词的平均相关词数量分布情况(本文中观察和分析的各种现象会受到不同语料库及计算方法的影响,但是只影响数值,不影响趋势)见表 1,随着阈值的减小,词条的平均相关词数量增加,对部分"明星"词条,如"教学"则拥有更多的相关词,远超过平均水平,膨胀更为严重。

表 1　相关词数量分布情况

阈值	平均相关词数量	"教学"相关词数量
0.9	2.88	88
0.8	3.61	165
0.7	4.54	329
0.6	5.95	648
0.5	8.11	1 315
0.4	11.11	2 486
0.3	16.37	4 791
0.2	20.54	7 762
0.1	24.12	11 128

1.2　文献保证挑战

无论具体策略如何,知识组织系统构建通常要考虑文献保证原则[20]。在具体的实践中,通常需要有文献语料库作为评估基础,但是大数据时代却面临着语料库不完备、不平衡和不准确的问题,这些问题在以前也存在,但是在大数据时代则变得更为突出。

1.2.1　不完备

由于计算分析能力的提升,用于知识组织系统构建的语料库规模逐渐增大,但是与需求之间的差距并未变小,仍不能完全覆盖需求。如某项实践中,基于特定检索策略提取约 2 000 万条二次文献数据作为语料库,在其中分析"移民浪潮"的相关词是"新加坡",但是没有"美国"这一典型移民国家,与常识不符,经过分析,实际上是语料库中并未包含足够的"移民浪潮"和"美国"共现情况。

此外,大量的语料来自网页抓取,而很多隐藏在数据库中的数据无法获取,这样不完备的语料库会影响基于其上的知识组织系统。

1.2.2　不平衡

真实世界本身是不平衡的,语料库尽管可以做一些筛选和调整,但是整体上仍然是不平衡的。根据词频分布的经验规律——齐普夫定律,词语本身在单篇长文章中的运用就是不平衡的,在语料库中同样也是不平衡的。由于某些词出现较少,大量根据统计相关方法获取的相关词就会更少。如在某实践中,按照同一标准计算,"移民浪潮"的相关词有 1 条,"高管"的相关词有 43 条,"中国特色社会主义理论"的相关词有 298 条,"教学"的相关词有 11 128 条。而依照先验知识库的方法同样受限于先验知识本身赋予者的局限性,依然无法避免不平衡性。

1.2.3　不准确

互联网环境下,知识传播速度更快,错误和偏离也会加速传播,如在科学技术领域常用词"阈值",在很多文献中被写作"阀值",如果依靠统计一般无法将其过滤。再如,在某实践中发现"高管"一词存在一个相关词"水墨画",且相关度达到 0.303,不符合常识,后来到语料库中检索发现,《艺术市场》刊物中有几期介绍一位名为"蔺高管"的画家的水墨画作品,但是在关键词中被错误地标注为"高管",并由于传播等原因,导致错误进一步蔓延。在大数据的资源条件下,如果不是人工核查,很难排除这种情况,而由于大数据体量和速度等因素,通过人工核查,也只能消除部分情况。

1.3　集成挑战

大数据环境下,需要尽可能通过集成等方式利用已经构建好的分散的知识组织系统,但是集成同样会带来很多问题。

1.3.1　概念定义不一致性

知识组织系统更关注概念,但是形式上还是体现为自然语言,在关联数据资源时,也往往依赖形式匹配。自然语言就存在一定程度的一词多义问题,尽管对专业领域选择术语的时候考虑了单义性,但是实际上很难保证,而在集成的时候,往往需要通过词形将不

同来源的知识集成起来，则就会存在与同一个词建立关联的词实际上来自不同领域的情况，进而将原来无关或者关联很小的词语通过较短的路径就能联系在一起，这势必会影响后续服务中的用户使用体验。如"信息生态学"一词实际上是由信息学和生态学学者分别提出来的，两种不同内涵的"信息生态学"尽管都是由信息科学和生态学交叉而形成的学科，但在研究对象、研究内容、研究方法等方面都存在着显著区别[21]。

由于大数据本身涉及的资源并非都是严谨的学术成果，因此各种缩写简写更为普遍，尤其英语首字母缩写，最终很难映射到知识组织系统的合适位置。如 IE 是一个常见的缩写，在不同的学科甚至同一学科代表不同的含义，它可能是 Industrial Engineering（工业工程）、Industrial Ecology（工业生态学）、Ionization Energy（电离能）、Information Extraction（信息抽取）、Information Element（信息元素）、Information Engineering（信息工程）和 Internet Explorer（微软的因特网浏览器）等。

根据 W3C 的简单知识组织系统（Simple Knowledge Organization System，SKOS）标准，在不同的概念体系之间映射，设置了 5 种映射或者对齐类型，分别为 skos:closeMatch（相似匹配）、skos:exactMatch（精确匹配）、skos:broadMatch（上位匹配）、skos:narrowMatch（下位匹配）和 skos:relatedMatch（相关匹配）[22]，这些映射类型相对简单，在实践上语义理解偏差会较大，要将不同的概念体系融合起来仍然存在极大的挑战。

1.3.2 关系定义不一致性

知识组织系统中一般都有一些关联关系，根据我国叙词表构建相关标准，大的类型包括等同关系、属分关系（等级关系）和相关关系。但是不同来源的知识组织系统之间关系的定义可能是不一样的，即使在标准中，也给出了关系类型的若干可能：等同关系包括同义词和准同义词，同义词有不同子类型，准同义词还可能包含反义词和部分事实上的属分关系；属分关系包含属种关系、整体－部分等级关系、实例关系；相关关系仅在标准中列举的就已经有 12 种，并且还不是完备列举[23－24]。所以形式上属于某种关系，实际上可能是不同的细分关系，尤其是在不同的知识组织系统中，这种表现更为明显。

1.3.3 构建方法不一致性

当前在知识组织系统构建方面，存在多种技术路线：完全依靠人工、依靠自动工具或者人机结合。早期大部分知识组织系统都是依靠人工构建的，近年来一些小规模的较为严谨的知识组织系统依赖人机结合，而一些较大规模的知识组织系统则主要依靠自动工具。自动工具一直在试图模仿人工，但是由于人工识别判定的内在机理不可能完全形成知识库或者被机器学习算法领会，因此自动处理总会有一些意想不到的结果出现。也就是说尽管很多自动工具的处理结果能够接近人工水平，但是总会有一些机器始终无法排除的结果，人工可以很简单地确定。因此出发点以及对质量的要求不同，构建方法不同，构建的结果自然也不同，把这些不同的知识组织系统整合在一起，自然也会存在不一致。

1.3.4 知识内容不一致性

知识组织系统本身会尽量避免不一致性，并且在一个较小的范围内也是可以做到的。但是，将多个知识组织系统集成为一个复杂的、广泛关联的适用范围更广的知识组织系统，则一定需要考虑容错，因为不一致的情况必然存在，而且也很难从全局进行协调。这就应该允许一定的非一致性，如同时承认鸟会飞以及鸵鸟、企鹅等少数鸟不会飞，这样可能并不会让服务变差，反而更接近于人的认知，当然要做好服务，还需要做好非协调推理，匹配不同知识内容的应用场景。对于有一些存在根本错误的知识，则需要根据用户的反馈或者抽检发现并及时修订。

1.4 应用挑战

1.4.1 需求多样性

知识服务的需求是多种多样的，不同的应用场景下，会用到不同的知识组织系统，如何将不同体量、不同深度的知识组织系统整合起来发挥作用，是应用中的一个难点，理想状态可能是构建一个大而全的知识组织系统，但是实践中受限于所需的人财物等资源很难实现。在应用多个知识组织系统的时候，要注意知识组织系统本身的覆盖范围和深度。例如，对于《中国图书馆分类法》这样的综合分类体系，很难覆盖像"燃料电池汽车"这样的细分类别，《汉语主题词表》这样的综合词表也不可能收录"粗晶粒钢"这样的专业词条并描述其与其他词之间的关系，更不要说收录"1,1,2,2,9,9,10,10－八氟[2.2]二聚对二甲苯"这样的复杂化合物名称。

1.4.2 外来适应性

在一些应用实践中，没有合适的知识组织系统，也很难从头构建，一种可能的方案是借助外来知识组织系统，但会带来外来适应性问题，如在某项实践中，没有合适的词表，将一个英文的金融银行领域词表翻译为中文，发现在翻译过程中有一些词很难翻译，如"10

-K”实际上是“(公司每年必须向美国证券交易委员会备案的财务报表等)公开文件”,但是不加以注释很难做出对应翻译,像401K(美国的一种退休金储蓄)这样美国特色的词条翻译过来,可能对中文数据组织没有多少意义,翻译后也会对一些关联关系造成影响,如原来具有关系“cash-synonym-money”,对词条翻译后关联变为“现金 - 同义词 - 现金”,失去了关联价值。此外,翻译后还存在两个词之间建立了多个不同关联关系的情况,这主要是因为两种语言无法做到一对一翻译。

2 应对策略

针对知识组织系统构建和应用的挑战,知识组织工作可以从标准化的可分结构与非对称结构、多来源更新策略和弹性应用服务模式等角度确立系列模型来部分应对。

2.1 知识结构

2.1.1 可分结构与标准化

面对大数据服务,知识组织系统不应也不能是唯一的,知识组织需要分工合作和共享。如在中国工程科技知识中心建设工作中,每个分中心根据自己的业务需求构建各自的专业领域知识组织系统,而知识中心从整体上整合各个分中心的知识组织系统,并加以补充完善,形成综合性的工程科技知识组织系统。可以假设理想的知识组织系统是一个大系统,某一个真实的知识组织系统是依据可分特性从大系统中提取出来的子系统,而这个特定的知识组织系统仍然可以继续具有可分特性。

据此提出大数据服务中知识组织系统的可分模型见图1,在知识结构上可分具体指“分层、分级、分块、分面”。“分层”是说从整体上知识组织系统全集可以分为频繁集和非频繁集两个部分。大数据服务总有一些在某个时间周期内频繁访问的知识组织系统数据,这部分就是频繁集,类似电商网站、新闻网站的热数据,与之相对的非频繁集则类似冷数据,这种分层及动态转化有助于应对用户对大数据服务的使用,在平均水平上改善用户体验。“分级”主要针对频繁集这部分,还可以进一步将其细分为核心集和扩展集,两者共同发挥作用,相对而言核心集更加稳定,对应长期不变的内容,而扩展集对应及时反应变化的内容。知识组织内容层次的流动往往发生在核心集与扩展集之间、扩展集与非频繁集之间,在具体的场景下,知识组织系统可以存在更多的层级,也可以对知识组织系统中的每条知识给出具体的稳定程度取值。“分块”是说知识组织系统的建设是分领域的,不同领域相当于全集的不同的块。“分面”是指即使对于一个领域,知识组织系统构建的目的和视角可能也是不一样的,同样对于新能源汽车这一分块,可能有的知识组织系统关注政策面,有的关注技术面,有的关注经济面。“分面”和“分块”对应着不同知识组织系统构建和应用群体的专业优势,因此对于知识组织系统内容质量提升帮助较大。

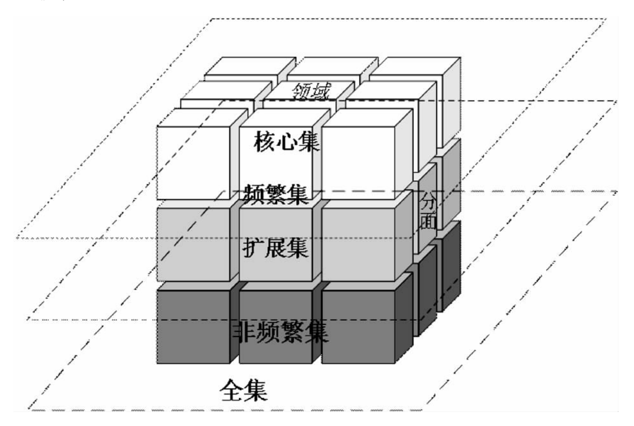


图1 大数据服务中知识组织系统的可分结构模型

可分结构可能带来潜在的集成风险,因此需要做好标准化工作。W3C提出的SKOS标准是知识组织系统在互联网上构建、共享和运用的事实标准,应用日益广泛。在叙词表领域国际标准化组织发布了ISO25964标准,我国也对应地更新了GB/T13190标准,在中国工程科技知识中心建设和新闻出版单位知识服务实践中也推出了相关的项目或行业标准,并推动相关标准进入国家标准。在分类法上,《杜威十进制分类法》《国际专利分类法》《中国图书馆分类法》《中国档案分类法》等分类法已经在大量的文献上使用,相关构建经验也可延伸到其他分类法的构建,同时,形成了一些标准,如《SDS/T 2121 -2004 数据分类与编码的基本原则与方法》,这些标准对于知识组织系统的规范化有一定帮助。当然,随着大数据自身的发展,标准化的程度还需要不断加强。

2.1.2 非对称结构

在叙词表构建标准中往往有成对关系类型对称指引的要求,而在关系逻辑校验中,往往将不对称作为校验的重要内容。但是在大数据时代,这种思路需要加以调整。实际上非对称指引早已有之,如分类法中的交替类目和正式类目,类似于叙词表中的用代关系,但是不在正式类目处标注[25]。现在只是将非对称指引扩大到原来严格要求对称指引的关系类型。

如在某实践中，按照某一标准，“教学”有超过1万条相关词，但是从对“教学”的认识角度出发，不太可能将1万余条词条同时展示给用户，也不太可能将这些词都加入到相关程序进行计算。在这些词中，根据相关程度排名靠前的“教师”“学生”“教室”“多媒体”等非常重要，排名靠后的“机械专业课”“体育游戏法”“写作知识”等相对于“教学”来说是相关的，但是并非“教学”的强关联知识。然而对于“机械专业课”“体育游戏法”“写作知识”这些词来说，和“教学”的关系则是较强的关联，因此从构建知识组织系统的角度没有必要因循对称性指引原则而为“教学”建立上万条关系，只需选择最重要的几条或者几十条即可，对其他的词条也同样处理，这就会出现“机械专业课”“体育游戏法”“写作知识”指向并关联“教学”，但是教学并不指向这些词，而是指向“教师”“学生”“教室”“多媒体”等词，即非对称指引。这种非对称结构实际上在微博这样的社交网络上更为常见，并且也被广泛接受，延伸到知识组织系统也非常合理。

2.2 多来源更新策略

知识组织系统的构建是不断迭代的，在大数据时代，完整更新一个知识组织系统是困难的，因此主要是面向可分结构中的某一局部进行修补式更新，具体更新模型见图2，更新的驱动力来自数据资源本身、用户和构建的应用本身。需要说明的是，这些更新往往是启发式的，通过发现关于某一知识的更新需求，往往需要考虑是否要将更新扩展到这一类知识上。

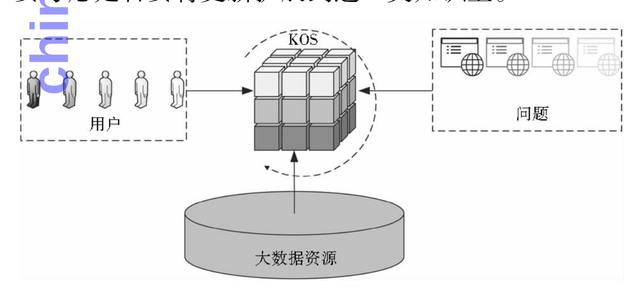


图2 大数据服务中知识组织系统的更新模型

2.2.1 资源驱动策略

资源长期持续变动是大数据的一个显著特征，并且变化的速度很快，因此为了组织变化的数据资源，对资源变动进行响应，需要根据资源修正和调整知识组织系统，尤其是变动频繁的扩展集部分，或者将一部分非频繁集变为频繁集，否则就会存在无法组织管理的数据，或者这部分数据揭示的水平不够。每当资源变化的时候，需要利用覆盖度等指标对已有的知识组织系统来评估，如果评估指标偏低，则需要更新相应的知识组织系统，或者确定该资源不符合知识服务的初衷而需要删除。

2.2.2 用户驱动策略

由于大数据的知识服务使用的数据和提取的知识不能保证百分之百是通过验证没有任何错误的，因此用户在实际使用中的表现可以作为发现问题和解决问题的一个途径。理想情况下，用户能够全面反馈使用中的问题，但是在现实中只能通过分析用户的行为，尤其是那些不关注、不点击、相对注意时间短等行为，挖掘出相关问题进一步改善，以便让用户在后续使用中不发生或者少发生这类错误。此外还应该关注系统对用户输入的响应，如果没有响应或者响应极少的情况，在排除用户输入错误的前提下，对应的知识组织系统很可能就需要调整。用户驱动实际上对用户数量以及用户数据处理能力都有较高的要求。如果用户数量少，则可能反映的问题没有典型性；如果数据处理能力弱，则必然存在更新时滞以及根本来不及更新的问题。

2.2.3 问题驱动策略

一项服务不能解决全部的问题，因此知识服务要相对专门化。解决的问题发生变化，对应的知识组织系统也需要更新，面向不同问题的知识组织系统基础可能是相似的，应用需要在这个基础上面向问题做定制化更新改造。比如金属材料方面的知识组织系统在面向下游机械行业科研人员设计的时候应该多考虑其性能指标相关力学属性，而在面向上游冶金行业用户的时候应该多考虑其晶体组织、冶炼工艺和冶炼设备之间的关系。做细分领域知识服务的时候，应根据问题逐步细化，只有在具体的“混合动力汽车”领域技术方向选择时才有必要区分“插电式”和“增程式”，而在更大的视角下，“混合动力汽车”本身甚至都没有单独列举的必要，而是代之以更高层级的“新能源汽车”。

2.3 弹性应用服务模式

借助知识组织系统提供大数据服务的机理见图3，知识组织系统和大数据资源一样，不是直接面对用户，而是通过针对不同问题的应用向用户提供服务。用户实际上受到大数据资源和知识组织系统本身的限制，大数据资源决定了他们能够获取服务数量的上限，当前的知识服务的竞争很大程度上是数据资源本身的竞争，因此各类知识服务的提供者倾向于获取并提供更多的资源，尤其是独占资源；而资源的标签、分类等知识组织系统的丰富程度和标引准确程度决定了服务质量的上限，在一些数据资源相对开放的领域，如专利、新闻，不同服务提供者能获取的资源是基本一致

的,因此竞争实际上变为包括知识组织在内的系列加工和服务技术的竞争,服务模式的弹性尤为重要。

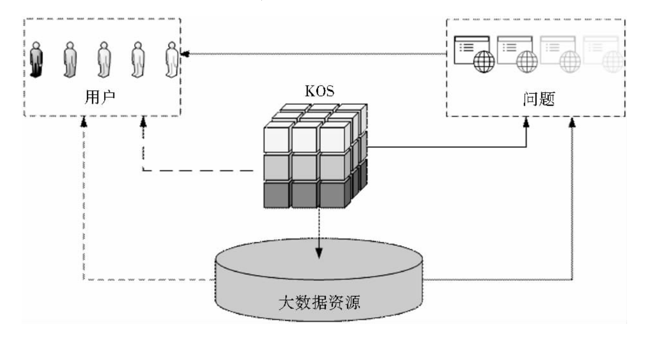


图3　借助知识组织系统的大数据服务机理

相同的大数据上,可以构建不同的应用服务,对应的知识组织系统及其关联的资源都存在体现不同维度、不同视角的一个或者多个分面,大数据服务中的知识组织系统特定应用模型见图4。不同的分面之间也没有明显的界限,某个分面具体用哪部分知识组织系统是根据资源、用户以及所面对问题动态调整的,知识组织系统应该是自适应弹性伸缩的,某些极限的情况可以包含全部的知识组织系统,但是一般会对知识组织系统的精细化程度要求有所降低,以节省计算和服务资源。

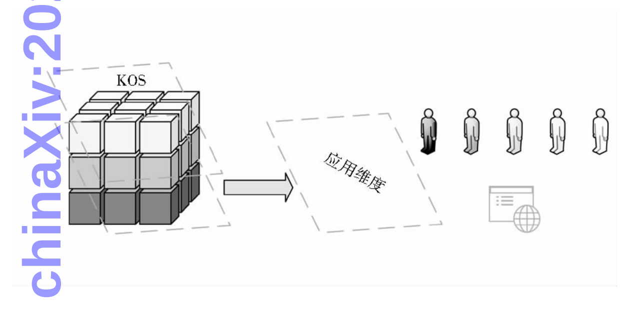


图4　大数据服务中的知识组织系统弹性应用服务模型

因为知识组织系统是迭代的,因此有多个版本,在利用这些知识组织系统的时候,要用相应的时间标签,标明在特定的知识组织系统状态下进行的各类标引操作。实际上一个应用中使用的知识组织系统的不同部分可能是相同时间的,也可能是不同时间的。如在专利标引工作中,用到国际专利分类表(International Patent Classification,IPC)标引发明专利和实用新型专利,用国际外观设计分类表标引外观专利,都需要通过标注版本号来区分同时使用的不同版本的分类表[26]。再如某档案大数据项目中,需要综合使用分类法和叙词表进行知识组织,其中分类法沿用1997年版本的《中国档案分类法》未做调整,词表则没有直接沿用1995年版本的《中国档案主题词表》,而是在其基础上补充了1995年后反映档案主题变化的新的主题词,如“一带一路”“亚投行”等。

3　结论

大数据服务中面临知识组织的数据膨胀挑战、文献保证挑战、集成挑战以及应用挑战。知识组织的膨胀挑战,可以利用标准化前提下可分的知识结构,通过多机构多人分工合作和共享来应对;文献保证挑战可以通过多来源的更新策略优化知识组织系统,并利用非对称性结构来应对;集成挑战可以通过标准化和可分机构来部分解决;应用挑战可以通过从可分的知识结构结合弹性服务模式和恰当的更新策略来部分解决。因此,在大数据时代,知识组织系统本身也需要与时俱进,形成一个可分层分级分块分面,非对称性,根据资源、用户、问题实时演化迭代的复杂系统。系统中可能包含若干相对独立但又相互关联的子系统,在应用中需要提取合适的子系统,与特有的数据资源、问题和用户结合形成服务,并在应用中更新知识组织系统。但是,在知识组织系统集成以及应用方面,挑战仍未得到彻底解决,其他对策中的真实案例也只是覆盖部分知识组织系统类型,还有一部分方案仍是设想,这些都需要下一步工作中逐步解决。

参考文献:

[1] FIDELIA I S, GEOFFREY C B. Implications of big data for knowledge organization [J]. Knowledge organization,2017, 44(3):187－198.
[2] MARCIO V, MARISTELAT H, EDISON I, et al. Transforming open data to linked open data using ontologies for information organization in big data environments of the Brazilian government: the Brazilian Database Government Open Linked Data - DBgoldbr[J]. Knowledge organization,2018, 45(6):443－466.
[3] 陈传夫,钱鸥,代钰珠.大数据时代的数字图书馆建设研究[J].图书情报工作,2014,58(7):40－45.
[4] 王曰芬,傅柱.大数据环境下知识表示与知识组织方法应用[J].数字图书馆论坛,2014 (3):32－43.
[5] 阳小兰,钱程,朱福喜.基于云计算的大数据服务资源评价方法[J].计算机科学,2018,45(5):295－299.
[6] 蒋永福. 论知识组织[J]. 图书情报工作, 2000,44(6): 5－10.
[7] 贺德方,乔晓东,朱礼军,等. 汉语科技词系统(新能源汽车卷)[M].北京:科学技术文献出版社,2012.
[8] 李旭晖,凡美慧.大数据中的知识关联[J].情报理论与实践,2019,42(2):68－73,107.
[9] 李旭晖,秦书倩,吴燕秋, 等.从计算角度看大规模数据中的知识组织[J].图书情报知识,2018(6):94－102.
[10] 孙坦,刘峥,崔运鹏, 等.融合知识组织与认知计算的新一代开

放知识服务架构探析[J]. 中国图书馆学报,2019,45(3):38 - 48.

[11] 陆泉,江超,陈静. 基于扩展疾病本体的电子病历大数据组织研究[J]. 图书情报知识,2019 (1):109 - 118.

[12] 潘刚,张运良,钟庆虹. 工程科技领域知识服务的思考与实践[J]. 情报工程,2018,4(5):4 - 12.

[13] 国家新闻出版广电总局数字出版司. 关于批准发布《知识服务标准体系表》等 8 项项目标准的通知[EB/OL]. [2019 - 04 - 30]. http://www. gapp. gov. cn/ztzzd/zdgzl/cbyszhzxsjxmzl/contents/4384/274644. shtml.

[14] 贺德方.《汉语主题词表》的回顾与展望[J]. 情报理论与实践,2010,33(2):1 - 4.

[15]《汉语主题词表(自然科学卷)》[M]. 北京:科学技术文献出版社,2018.

[16] TSINGHUA UNIVERSITY KEG. XLORE[EB/OL]. [2019 - 04 - 30]https://xlore. org/.

[17] NATIONAL LIBRARY OF MEDICINE (US). UMLS reference manual [EB/OL]. [2019 - 04 - 30]. https://www. ncbi. nlm. nih. gov/books/NBK9679/.

[18] 中国工程科技知识中心. 中国工程知识中心词表总表核心集 CKT-C [EB/OL]. [2019 - 08 - 09]. http://data. ckcest. cn/filedata/index. html? p = detailDoc&i = 8a9a2dd466539c2d01665bc34c510000.

[19] DOUGLAS L. Cyc[EB/OL]. [2019 - 04 - 30]. https://en. wikipedia. org/wiki/Cyc.

[20] 刘菡. 略论信息组织中的文献保证原则[J]. 国家图书馆学刊,2019,28(1):57 - 65.

[21] 娄策群, 桂晓苗, 杨小溪. 我国信息生态学学科建设构想[J]. 情报科学, 2013, 31(2):13 - 18.

[22] W3C WORKING GROUP. SKOS simple knowledge organization system primer[EB/OL]. [2019 - 04 - 30]. https://www. w3. org/TR/skos-primer/.

[23] 全国文献工作标准化技术委员会. 汉语叙词表编制规则:GB13190 - 91[S]. 北京:中国标准出版社,1992.

[24] 全国文献工作标准化技术委员会. 信息与文献叙词表及与其他词表的互操作第 1 部分:用于信息检索的叙词表:GB/T 13190.1 - 2015[S]. 北京:中国标准出版社,2015.

[25] 王小华.《中图法》第 5 版交替类目研究综述[J]. 图书馆学刊,2015,37(11):132 - 134.

[26] 中华人民共和国国家知识产权局专利局. 中国专利文献著录项目: ZC0009 - 2012 [EB/OL]. [2019 - 08 - 09]. http://www. sipo. gov. cn/wxfw/zlwxxxggfw/zsyd/bzyfl/zlwxyxxbz_gnbz/1053740. htm.

The Challenges and Countermeasures of Knowledge Organization in Big Data Service

Zhang Yunliang[1,2]

[1] Institute of Scientific & Technical Information of China, Beijing 100038

[2] Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, Beijing 100038

Abstract: [**Purpose/significance**] Big data services bring more challenges to knowledge organization. Through observing, understanding and analyzing these challenges, knowledge organization work stakeholders would grasp possible changes and provide countermeasures. [**Method/process**] Focusing on the construction and application of knowledge organization system, challenges of different aspects of knowledge organization were analyzed and countermeasures were proposed from related real case practice. [**Result/conclusion**] The challenges of knowledge organization in big data services can be divided into four aspects: data explosion, document assurance, integration and application. A series of knowledge organization frameworks for big data services including new knowledge structure, multi-source updating strategy and elastic application service model are proposed, which can better meet the above challenges.

Keywords: knowledge organization big data knowledge service challenge countermeasure